DOCUMENT RESUME

ED 199 275 TM 810 176

AUTHOR Echternacht, Gary
TITLE The Uses and Misuses of Test Scores: Technical Assistance Perspective.
PUB DATE Apr 81
NOTE 24p.: Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; Compensatory Education; Elementary Secondary Education; *Program Evaluation; *Scores; *Standardized Tests; *Testing Problems; *Test Interpretation; Test Results
IDENTIFIERS *Elementary Secondary Education Act Title I; Technical Assistance Centers; *Test Use

ABSTRACT

The uses and misuses of standardized test results used for program evaluation as seen by a staff member of an Elementary Secondary Education Act (ESEA) Title I Technical Assistance Center are described. In ESEA Title I, test scores are used to select students for the program. Although federal requirements do not require using standardized test scores for selecting students, most states encourage or require their use. Test scores are also used by both regular classroom teachers and Title I teachers to form ability groups for instruction. Frequently a school will either administer a diagnostic test to Title I students or use standardized test results to make general diagnoses. Test results are also used for short term and long term program evaluation. Four categories of the misuse of standardized test results are: (1) "filed and forgotten" which indicates that the test scores are used only for mandatory reporting; (2) "big bang" which criticizes the overinterpretation of test scores; (3) "reliance on grade equivalent scores" which confuse interpretation; and (4) "interpreting summative evaluation results without formative evaluation results" which represents an explanation without a proper context. (Author/RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED199275

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

The uses and misuses of test scores: technical assistance perspective

Gary Echternacht
Educational Testing Service

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G. Echternacht

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 810 176

Paper presented at the American Educational Research Association's Annual Meeting, Los Angeles, 1981.

Abstract

This paper describes the uses and misuses of the standardized test results used for program evaluation as seen by the ESEA Title I Technical Assistance Center staff member. The paper is neither a review of literature nor a detailed analysis of the issues associated with test use, but rather a record of observation made in the field. Four general misuses are identified. "Filed and forgotten" indicates that the test scores are used only for mandatory reporting. The "big bang" is a criticism of the overinterpretation of test scores. Grade equivalent scores are criticized as well as the failure to link norm-referenced test results to unit or instructional test results.

The uses and misuses of test scores: technical assistance perspective

Regional Technical Assistance Centers were created by the federal government in 1976 to help both states and local school districts comply with ESEA Title I regulations on evaluation and to improve the quality and usefulness of those evaluations. Rather than act as regulators, the centers were established as houses of technical expertise, where people could be called on for assistance in testing and evaluation matters. The centers provide formal training and informal consulting services to states and local school districts within their geographic regions. Although standardized norm-referenced tests are not mandated directly by Title I regulations, they are indirectly, and indeed they form the cornerstone of the required evaluation. Local school districts must report their evaluation results using a scale that is based on a national norms population. This requirement has resulted in a shift by local districts away from criterion-referenced testing and toward norm-referenced testing for outcome evaluation.

With regard to test use, technical assistance centers are most concerned with the tests used for outcome evaluation. In the typical Title I evaluation, the worth of a Title I program is judged in terms of pretest-posttest gains in percentile rank. Schools use the test results for more than just program evaluation, however, and the centers often become involved in helping schools use these tests for other purposes. Staff from the centers are well trained in measurement and evaluation techniques and usually are the only such experts who routinely contact school people.

The purpose of this paper is to represent some of the uses and misuses of standardized test scores from the perspective of a technical assistance staff member. The paper is based on my observations as a technical assistance center staff member for the past four years and is neither a review of the literature nor a particular incisive examination of issues in how tests are used. It is directed toward standardized achievement tests that are used for program evaluation.

Testing and Title I Evaluation

The generic model our center uses for evaluation is given in Figure 1. Information comes from test scores, questionnaire results, student records, and observations. Communication is the transmitting of the information through various written reports, visual displays, and oral presentations. Decision is the process used for assimilating the information into some policy statement.

In our model of evaluation we adopt Phillips' (1980) distinction between facts and values in the process of making policy. The information gathered by the evaluator is factual—i.e., "scores were higher at Grant School than they were at Lee School," "the gain in fourth grade was 4.2 NCEs," "teachers at Lee School did not use the ABC reading program." The role of the evaluator is to collect, assemble, and present this factual information for the managers and policy makers who are ultimately responsible for using, or not using, the information. Although the evaluator may be part of a decision making group or in some instances may be the decision

maker, the responsibility for decison making lies with the manager or policy maker. As Phillips points out, values come into play during the decison making process because values are needed to link the information with the policy. It is the questioning of the values of some educators that is at the heart of my criticisms of the way in which test results are used.

The general model for evaluation does not specify who is responsible for making decisions. In some instances the evaluator is also the program manager and is responsible for selecting students, diagnosing their needs, and placing them in appropriate groups in addition to evaluating the program. In other instances, the program managers do not act as evaluators. Sometimes high level administrators are involved in making programatic decisions. The way in which test and evaluation results are used, I suspect, is influenced by who and how many decision makers are involved. As more people at higher levels are involved in decision making, the greater is the chance that nonevaluation factors can dominate the decision making process. For example,

- evaluations with negative implications are suppressed while positive evaluations are overemphasized.
- evaluations of programs relying on patronage appointments (e.g., classroom aides) must be positive regardless of the evaluator's findings.
- newspapers require either sensational or highly significant results, so negative results are suppressed and positive

results are reported to bolster the standing of the superintendent.

Typical Uses of Test Scores

Contrary to the impressions one might receive (for example, see David, 1978) standardized test scores are used by schools. As almost any introductory textbook on measurement will point out, tests are used for selection, classification, placement, and evaluation. Sure enough, tests are used for these tasks although usually in conjunction with other measures.

In ESEA Title I test scores are used to select students for the program. Although federal requirements do not require using standardized test scores for selecting students, most states encourage or require their use. Program managers are responsible for selecting students.

Plans for selecting students vary from the simple to the exotic. In a few schools students are selected for the program based only on their test score. A cut score is set and strictly enforced. More often, teacher judgment is allowed to enter the process. Sometimes this takes the form of a rating scale where each student's achievement is rated by his teacher. Sometimes the ratings are unsystematic and not well defined. Test scores and teacher ratings may be combined into a composite score, with selection based strictly on the composite. Or, schools may set up multiple cut-offs on the test scores and teacher ratings, requiring that the student score below both or in some instances below either cut-off to be selected for the program.

In selecting students, we find two problems occurring. In the first, test scores are not involved. The problem here lies in the way teacher ratings are obtained. Often the method is neither systematic nor objective. Constructs are not well defined. Procedures for responding are ill defined. Training teachers in how to rate students and complete the rating forms is minimal. The result is unreliable and invalid ratings. The second problem involves how to set the cut scores. The number of students that can be served is constrained by the type of curriculum offered and the funds available. Therefore, cut scores are set by trial and error in any manner that fits the constraint. Because the cut score setting process is complex, little consideration as to the appropriateness of the result is given. Validation of the selection process, establishing the validity of the variables and selection rules, is rare.

Test scores are also used by both regular classroom teachers and Title I teachers to form ability groups for instruction. Unless a student is new to the teacher, the test scores are moderated by a heavy dose of teacher judgment. This judgment is based on past observation, often from the student's past teacher. For the new student, test scores are often the only information available to place students into ability groups at the beginning of the year. The placement procedure is usually left up to the teacher, involving a few trials before a satisfactory result is achieved. Placement is an important function for test scores in Title I because teachers provide instruction in small groups and they want to work with homogeneous groups.

Title I regulations require that student instructional needs be diagnosed prior to writing instructional objectives for the program. Although a school need not use tests to make diagnoses, most frequently a school will either administer a diagnostic test to Title I students or use standardized test results to make general diagnoses. Some standardized tests encourage such diagnosis. For example, both the California Achievement Test and the Metropolitan Achievement Test, Instructional Form, provide objective scores for the various objectives tested. Teacher judgment is also used to diagnose need although it is not used to the same extent as in selection and placement.

As noted earlier, standardized achievement tests are used for program evaluation. They are used to indicate the short term (one year) impact of the program. Program impact is defined in the Title I evaluation system as the difference between end of year standing on an achievement test and a predicted level of achievement. The predicted achievement level is a statistical prediction based on previous standardized test results. This approach is prescribed in the Title I evaluation regulations, and most school districts apply the methods without significant difficulty.

Test results are also used for longer term evaluation. Exhibit 1 shows the results of one such long term evaluation. In this example, test scores for a cohort of students were assembled over two years. Standardized test scores provide a two year picture of the effect of the program. In this example one can see that the two year effects are not

the sums of the short term effects. Such evaluations have been stimulated by new evaluation requirements that schools must evaluate the "sustained effects," or long term impact of their programs at least once each three years.

The effects of Title I evaluations have been manifold in my judgment. They have stimulated closer examination of district testing programs. Schools are more interested in having their testing programs measure relevant curriculum objectives. The accuracy of the test scores has improved because of more emphasis in appropriate test administration. Program managers and teachers are more interested in motivating students to take the tests so their results will be more valid. Test scores are being used to judge the appropriateness of the curriculum, not as the only indicator, but as a performance measure on a sample of objectives that are believed "nationally representative." Formative or process evaluations are beginning to occur explaining the results of the summative evaluations. These effects have occurred because the Title I evaluation system is essentially normative. A school can compare its results with other schools in the same state or with other schools in other states.

Misuses of Test Scores

As said earlier, any statement about misuses of test scores implies some type of values statement. Before listing any misuses, let me list some of my beliefs with regard to testing and evaluation.

1. Standardized achievement test results are generally valid achievement indicators and should be used for selection, placement, classification and evaluation.

2. Standardized achievement test results are not absolutely valid and should not be used in isolation of other valid information.

3. Standardized achievement test results should not be used to place students in a linear, one-dimensional, lock-step developmental continuum.

4. Standardized achievement test results and evaluation results should be linked to performance on instructional units.

In the following, I give four examples of misuse based on the above statements.

Filed and forgotten. With regard to evaluation findings, this type of behavior is the theme of David's (1978) report on the uses made of local Title I evaluations. In her report, she asserts that the extent to which evaluations are used is not dependent on the design, quality, or technical merit of the evaluation. Instead they are used to substantiate a priori positions. If the evaluation is positive, it is offered as scientific evidence of effectiveness; if the evaluation is negative, it is filed and forgotten.

As another example of filed and forgotten, let me relate a personal incident. After my eldest child took her first standardized achievement test, my wife and I were anxious to see the results. For two months we waited for the score report to come home. Summer vacation came. No results came home. When school began in the fall, my wife, who visited the school regularly, asked our child's former teacher if we could see our child's scores. The teacher said they were never sent to the teachers and suggested speaking with the principal. The principal said they were never sent to the schools, but she would see to it that we would get a copy of the scores. Weekly, my wife asked the principal about the scores. About four weeks later we received a xerox copy of a score report for our child. Only our child's scores were sent to the school. The others remained in the central office.

As it turned out, the test scores were returned to the district from the scoring service four weeks after the testing. Once they were returned, the central office staff began working feverishly on the evaluation report that was to be submitted to the state. This had the testing office's highest priority and all else stopped so the report would be finished before summer. After the scores were transcribed and used to select next year's participants, they were filed and over the summer forgotten.

Although such action is often a result of poor planning or management, underneath it lies a fundamental belief. It asserts that test scores are not valid and should not be used for anything other than filling out the required forms and reports. It is reflected in a preference for observational

and judgmental evaluations over evaluations using tests. It is characteristic of people who "know their programs are good" and by many teachers who see tests as being a negative evaluation device for students who need positive educational experiences.

The big bang. This type of behavior is more a fear than a reality. I know of only one dramatic example although in that instance the effect went well beyond the school district and the incident is remembered clearly even though it happened nearly four years ago. As a result of published state administered test results, the mayor of one city, a city whose schools scored at the bottom of the list, fired all his appointed school board members because the tests indicated that "the schools were doing a poor job." The decision was swift, dramatic, and without appeal.

On a more modest scale, the big bang often occurs when students are selected for Title I programs. In some instances only a test score is used and it is strictly adhered to. There is no appeal or room for other information. Student A has more ability than student B because his test score is higher, regardless of the difference in test scores. This policy is often made part of state regulations for educational programs and is interrelated with minimum competency testing.

Using a test score as the only criterion for selection has several disadvantages. For example, test scores are most unreliable for selecting those students who score near the cut-off. Repeated testing could easily

result in alternately scoring above and below the cut-off. In such situations it is difficult to say with assuredness whether or not the student belongs in the compensatory education group. Selecting students is also an individual decision while evaluation is a group decision. The effect of this distinction is that individual decisions require more data than group decisions to achieve the same degree of accuracy. Therefore, we want to use more information to select students than we might use to evaluate a program. If we use the notions of item response theory to conclude that test reliabilities vary according to ability, for the typical standardized test we conclude that scores are most unreliable for those students who are exceptionally low or high in ability. Standardized tests also sample limited content areas in limited test administration times. All are reasons for not using test scores alone to select students for compensatory education programs.

Behind the big bang is the belief that test scores are completely valid and can and should be used in isolation of other information. People subscribing to the big bang distrust subjective or clinical assessments or observational ratings. They tend to be administrators or policy makers who must deal with a public clamor for more accountability or responsibility. This view is not widely held by those within the educational system.

Reliance on grade equivalent scores. Unlike the other misuses I point to, this one can be attacked on both factual and value grounds. Grade equivalent

scores were intended to help interpret test scores by providing a reference that most people understand, showing progress through grade levels. Unfortunately, grade equivalent scores only confuse interpretation. The grade equivalent scale implies existence of a linear equal interval scale. When we look more closely, however, we begin to see deficiencies in this approach. For example, if a student whose age places him in the second grade comes to a school, is given a standardized test and if he is found to score "one grade below level" on one widely used reading test, we will find that this score falls below the first percentile of the national norms or more than two and a half standard deviations below the mean. Most probably that student is a low achiever who would benefit from some compensatory program. For a second student who is in the eighth grade and who also scores "one grade below level," we will find this score falling at the 34th percentile or about two fifths of a standard deviation below the mean. Such a student could easily be called average in achievement, with abilities similar to other eighth graders.

Grade equivalent scores are supposed to indicate "the grade level at which the typical pupil makes this raw score (Iowa Tests of Basic Skills, Teachers Guide, 1978, p. 23)." If a student makes a grade equivalent score of 4.3 it is supposed to mean that the raw score on the test is the same as that made by the typical or median student in the fourth grade at the end of the third month. But if we look at the students in the norms group, we find no students at the third month in the fourth grade ever took the test. With the exception of the median scores at

the time students took the test, all other grade equivalents are extrapolated, interpolated, or nonlinear transformations of other scores.

Grade equivalent scores, or grade level norms, are also taken as a performance standard. I have always been amazed by people reacting to the statement that half the students in our schools are below grade level. To much of the public, to be at grade level is a minimum standard, not just the median score in the national norms.

Evaluators and many school people are skeptical of grade equivalent scores, but they are often forced to report them to administrators and the public who believe in them. Many a teacher or program director after explaining the meaning of a student's score in terms of percentile rank or standard scores is asked, "At what grade level did Mary score?" Herein lies the inappropriate value. People believe a norm referenced test should place a student on a linear, one dimensional continuum showing progress through a lock step sequence of instruction. The standards used to place students on this continuum will be performance standards, where at grade level is equivalent to mastery. It is this value that leads people to conclude that a student either should or could be placed in the grade at which he scored.

Interpreting summative evaluation results without formative evaluation results. Standardized tests are used for summative evaluation in the Title I system. Among the states and technical assistance centers,

there has been an emphasis placed on using evaluation results to improve programs. States have begun to develop norms for evaluation results so schools can compare their results with those of other schools. This helps schools better interpret and use their evaluations.

We should not fall into the trap of thinking that one evaluation tells the whole story. If we were to perform the evaluation again, our results would differ. That is a fundamental principle of statistical analysis. The variation in results may not be large and will depend on test administration practice, sampling fluctuations, and the skills of those responsible for processing the data. As evaluations are repeated, however, we gain more confidence in their results and begin to believe the picture being painted.

Overreliance on summative evaluation results can be traced to two factors. First, with the current emphasis on using evaluation results, people are going to try and use the information they have. For most schools, this means that only summative evaluation data will be available. Second, many people want to look at only "the bottom line," which usually means the summative results. Formative or process evaluation is considered an extra because only summative evaluation is believed important.

Summative evaluation is important. It is probably a very good "bottom line," but the results of a summative evaluation should be linked to the results of a formative or process evaluation. A good program design consists of a series of units, the accomplishments of which results in

overall improvement in achievement. A good formative or process evaluation will assess the extent to which the units were successfully completed. A good overall evaluation will be able to link the summative results with the formative results dealing with the instructional units.

The analogy with football games is good. The summative results or the "bottom line" is the final score of the game. Nevertheless, no coach uses only the final score as a way to evaluate his team's performance. He studies game films, key plays, and the quality of the opponents when evaluating his team's performance. This is process evaluation, and the football coach links the process, the plays comprising the game, to the outcome as a way to explain what happened. So too will the good evaluator link the performance on instructional units to performance on the test.

Judging from the winter issue of Using Evaluation Results, a recent newsletter published by the Center for the Study of Evaluation, formative evaluation is being promoted among evaluators. The theme of most authors has been that formative evaluation should be completed and the program running as designed before summative evaluation begins. In other words, formative evaluation is viewed as a prerequisite for summative evaluation.

In my own observation this is not the way evaluations have been performed in the past, nor is it realistic to assume that this model will work in the future. We in the technical assistance centers urge schools to conduct formative evaluation. We have specific training programs aimed

at showing school people just how to conduct such evaluations. It is interesting to note, however, the reasons school people give for wanting to start formative evaluations. They want to obtain data that will explain the summative evaluation results. If the evaluation is negative, schools want to be able to explain the findings in terms of defiencies in the way the program was run or in terms of contextual events that prevented the program from working the way it should. If the evaluation is positive, schools want to know what elements of the program contributed to the success. Regardless of how the link is made, it is clear that the motivation for formative evaluation has come from summative evaluation. Without summative evaluation, I doubt if formative evaluation would much occur.

Summary

From the vantage point of the technical assistance center, I have seen that standardized test results are used. Their uses include for selection, classification, placement, and program evaluation. In some instances, tests are misused in these contexts when they are used exclusively and other information is not taken into account. Typically, test scores are used in conjunction with other information to reach decisions. Often decisions are based on many years of experience. Because of this, test results do not always have dramatic and obvious consequences.

Nevertheless, test results are misused. When they are "filed and forgotten," information that is valid and reliable is lost. In what I

have called the "big bang" effect, test results are used at the exclusion of other information. Attempting to generalize from grade equivalent scores has led people into misrepresentations. Finally, explaining evaluation results without a proper context represents a misuse similar to the "big bang" effect.

What can be done to make test use more informative and to reduce misuse? School administrators and teachers must be educated in principles and effects of educational measurement. Graduate and undergraduate education schools must emphasize the notions of measurement error, test validation, and score interpretation, as well as how to use these notions in decision making. Then too, professional evaluators must bear some of the responsibility for past misuses. Too often we have tried to veil evaluation in a cloak of methodology and jargon. Our reports can become more clear. Our presentations can be more geared to the interests and expectations of our audiences. We can be more concerned with providing results in an effective manner that are important to the decision maker.

References

David, J. Local uses of Title I evaluations. Menlo Park: SRI International, July, 1980.

Iowa Test of Basic Skills, Teachers Guide. Boston: Houghton Mifflin, 1978.

Phillips, D. What do the researcher and the practitioner have to offer each other. Educational Researcher, 1980, 9, 11, 17-24.

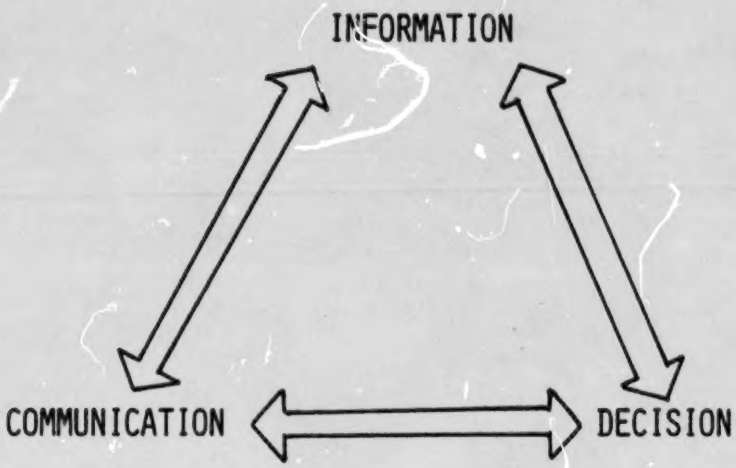
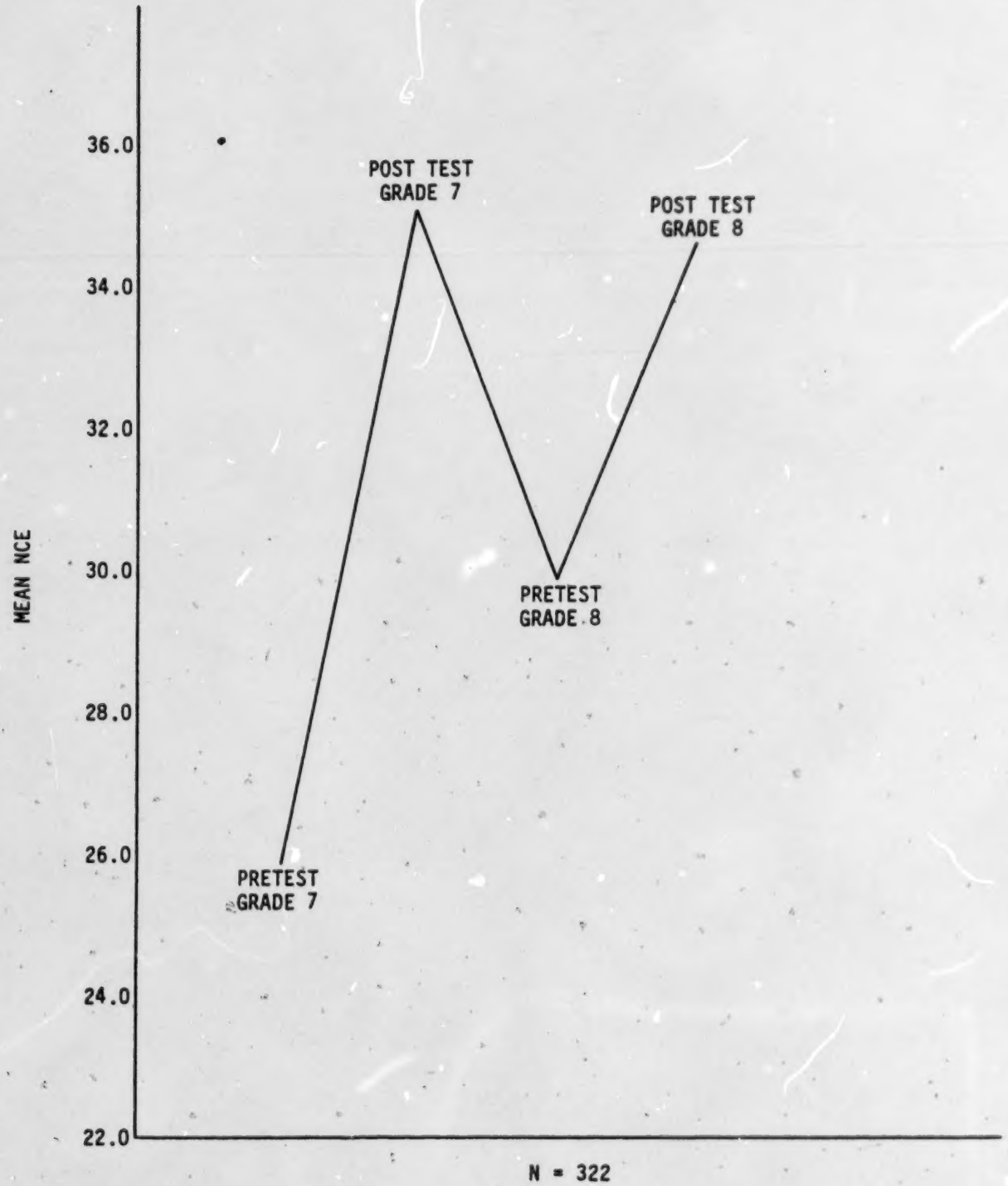


Figure 1. Technical Assistance Center model for program evaluation.

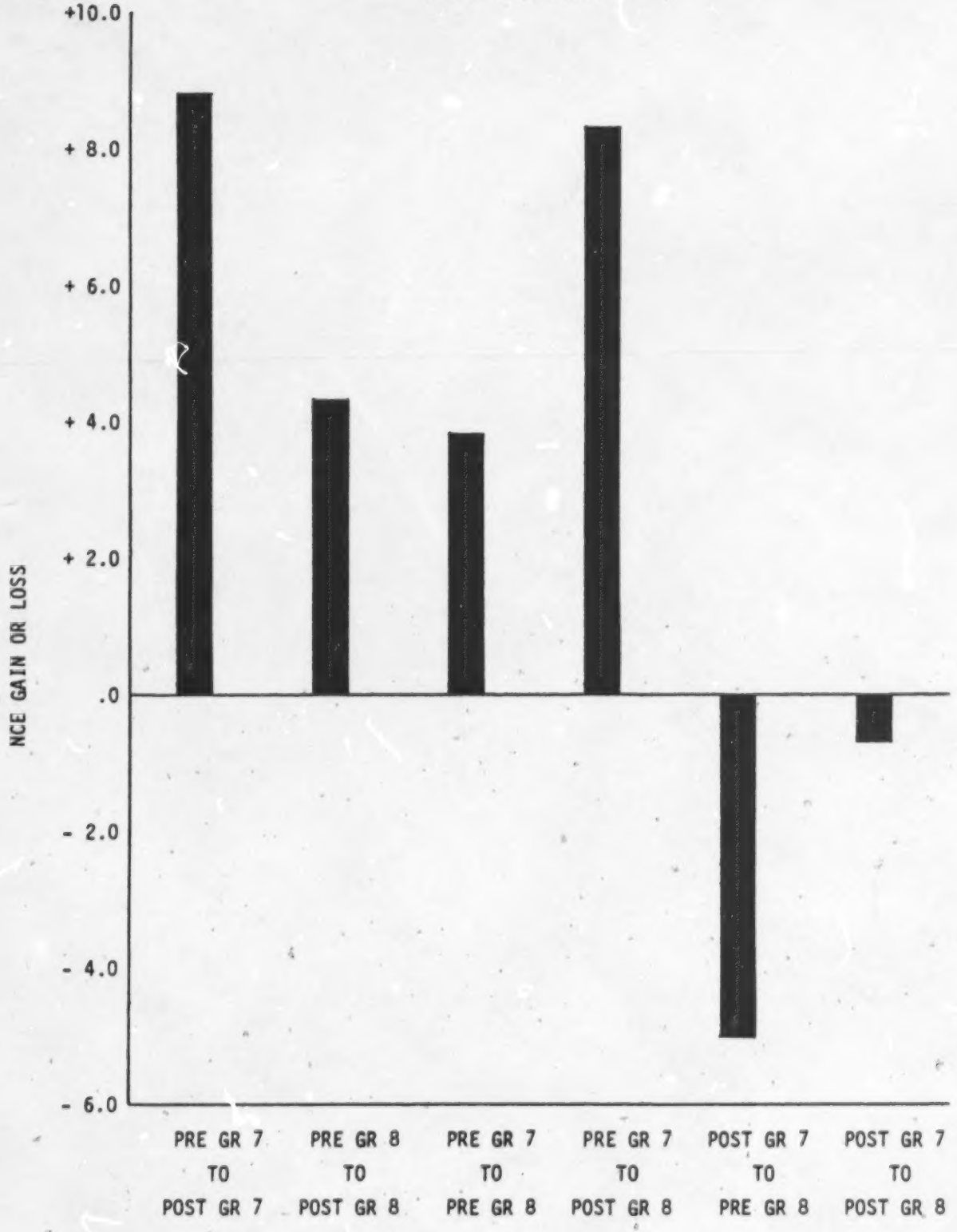
SUSTAINING EFFECT OF TITLE I

CAT READING SCORES FOR GRADE 7 PUPILS WHO REPEATED PROJECT IN GRADE 8

1978-79 to 1979-80



SUSTAINING EFFECT OF TITLE I
CAT READING SCORES FOR GRADE 7 PUPILS WHO REPEATED PROJECT IN GRADE 8
1978-79 to 1979-80



N = 322